Beyond complete genomes: from sequence to structure and function

Eugene V Koonin*, Roman L Tatusov and Michael Y Galperin

Computer analysis of complete prokaryotic genomes shows that microbial proteins are in general highly conserved – ~70% of them contain ancient conserved regions. This allows us to delineate families of orthologs across a wide phylogenetic range and, in many cases, predict protein functions with considerable precision. Sequence database searches using newly developed, sensitive algorithms result in the unification of such orthologous families into larger superfamilies sharing common sequence motifs. For many of these superfamilies, prediction of the structural fold and specific amino acid residues involved in enzymatic catalysis is possible. Taken together, sequence and structure comparisons provide a powerful methodology that can successfully complement traditional experimental approaches.

Addresses
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA
*e-mail: koonin@ncbi.nlm.nih.gov
Correspondence: Eugene V Koonin

Current Opinion in Structural Biology 1998, 8:355–363

http://biomednet.com/elecref/0959440X00800355

Abbreviations
COGs clusters of orthologous groups
HAD haloacid dehalogenase

Introduction

The determination of the complete genome sequences of several bacteria and archea and one eukaryote [1–6,7••–12••] marked the beginning of a new age in biology. For the first time, we can take a look at the complete set of proteins present in the cells of each particular organism and try to identify the proteins responsible for each cellular function. In cases where no known proteins can be found to perform a particular task, the most likely substitutes can be predicted from the set of unassigned gene products. Clearly this can be done only by analysis of complete genomes, as partial sequences do not allow us to ascertain that certain proteins are not encoded in a given genome [13]. These new approaches are gradually changing our understanding of a variety of biological phenomena. As the number of sequenced genomes is expected to grow exponentially for the next few years, their impact on different biological disciplines will increase. We have recently discussed the implications of the complete genomes for microbial evolution [14]. Here we consider the effect of the genome revolution, together with the improving methods for sequence analysis, on our ability to predict and understand protein structure and function.

Towards a natural taxonomy of proteins and protein families

The numerous genome sequencing projects have resulted in a rapid growth of protein databases (see, e.g. [15]). In contrast to the pre-genome era, when researchers typically chose to clone and sequence genes with documented functional roles, we are now getting many protein sequences whose functions are not known. This presents a challenge to extract the most from these sequences in terms of salient features of the encoded proteins, for example to classify them according to their homologous relationships, and to predict their possible catalytic activities and/or cellular functions, three-dimensional (3D) structures and evolutionary origin.

Protein classifications, pioneered by Dayhoff and her coworkers, have historically been based on sequence alignments. Similar proteins formed families, which were combined into superfamilies [16]. This approach, continued in the PIR database [17], proved extremely popular. However, even PIR superfamilies often unite closely related proteins and more distant relationships are being missed. Other protein databases, such as PROSITE [18], PRINTS [19], Pfam [20], and ProDom [21], group proteins on the basis of conserved sequence motifs and, generally, contain much more diverse protein families. Structural comparisons of proteins, implemented in FSSP, CATH and SCOP databases, offer yet another approach to protein classification [22–24]. SCOP superfamilies, for example, unite proteins that have some similarities in their 3D structures, but often no detectable sequence similarity [25]. Thus, in the absence of clear sequence or structural similarities, the criteria for inclusion of distantly related proteins into a family (or superfamily) become increasingly arbitrary.

With the inception of extensive genome sequencing, it has become possible to classify genes and proteins on a different principle, namely by delineating families of paralogs — related genes within the same genome [26,27]. Such analyses have revealed a complex hierarchical organization of paralogous families in each of the studied genomes and produced at least two generalizations: first, the fraction of genes that belong to families of paralogs increases with the increase of the total number of genes in a genome: from ~25% in the minimal genome of *Mycoplasma genitalium* to >50% in the large (for a prokaryote) *Escherichia coli* genome; second, the largest superfamilies of paralogs are mostly the same in all genomes [28–33].

Knowledge of all the protein sequences from multiple complete genomes (Table 1) allows us to redefine the entire

Table 1

Protein families and 3D structures in complete genomes.

Species	Proteins encoded in the genome*		COGs found (% total)	3D structures	
	Total number	Belong to COGs† (% total)		In PDB	Predicted‡
Escherichia coli	4289	2003 (47%)	821 (95%)	240	667
Haemophilus influenzae	1717	979 (57%)	658 (77%)	2	267
Helicobacter pylori	1566	841 (54%)	617 (72%)	0	169
Synechocystis sp.	3169	1551 (49%)	796 (93%)	2	431
Borrelia burgdorferi	850	483 (57%)	363 (42%)	0	105
Bacillus subtilis	4100	1945 (47%)	732 (85%)	12	578
Mycoplasma genitalium	467	341 (75%)	290 (34%)	0	75/103
Mycoplasma pneumoniae	677	378 (56%)	309 (36%)	0	78
Methanococcus jannaschii	1715	830 (48%)	498 (58%)	0	170
Methanobacterium thermoautotrophicum	1869	897 (48%)	484 (56%)	0	199
Archaeoglobus fulgidus	2407	1131 (47%)	512 (60%)	0	290
Saccharomyces cerevisiae	5932	1736 (29%)	577 (67%)	45	846
Caenorhabditis elegans	12,178	2172 (18%)	466 (54%)	2	NA

*The numbers are from the latest updates in the GenBank genome division (ftp://ncbi.nlm.nih.gov/genbank/genomes). *C. elegans* genome is about 85% complete; the data are from Wormpep12 (www.sanger.ac.uk/Projects/C_elegans/wormpep). †Based on the set of 860 COGs, obtained by adding *H. pylori* proteins to the original set of 720 COGs [37••]. ‡The numbers are from the PEDANT database [53•], calculated by comparing the protein set encoded in each genome to the PDB using FASTA with cutoff score of 120; the second figure for *M. genitalium* is from [54•]; the data for *C. elegans* are not available.

problem of protein classification. Since the fraction of proteins conserved over large phylogenetic distances (ancient conserved domains) appears to be nearly constant at ~70% in all prokaryotic genomes [34•], it becomes feasible to replace more or less arbitrary clustering of proteins by similarity with consistent groups in which the evolutionary relationships between the members are specifically defined. Such a classification of proteins can provide a framework for evolutionary studies and for rapid, largely automatic, functional annotation of newly sequenced genomes.

Several classifications of homologous proteins encoded in complete genomes have been produced, based on all-against-all protein sequence comparisons [35,36,37••]. Each of these projects is aimed at the identification of orthologs, that is direct counterparts in different genomes, connected by an uninterrupted line of vertical descent and typically retaining their physiological function [26,27]. In particular, the system of clusters of orthologous groups (COGs) was designed to accommodate the vastly different evolution rates observed for different genes [37••]. The COGs construction procedure identifies the closest homologs in each of the sequenced genomes for each protein, even if the similarity is fairly low and not statistically significant by itself. The approach to the identification of COGs was built upon the transitivity of orthologous relationships, that is the simple notion that any group of at least three genes from distant genomes, which are more similar to each other than they are to any other genes from the same genomes, is most likely to belong to an orthologous family. Clearly, this is a probabilistic assumption based on a 'weak molecular clock concept', which posits that orthologs are more similar to each other than they are to paralogs with different, even if related, functions. This assumption, however, seems to hold true in cases where we have reasons to accept orthology on functional grounds (for example, aminoacyl-tRNA synthetases or ribosomal proteins). Orthology is not necessarily a one-to-one relationship, as in cases of lineage-specific duplications, orthology can only be established between families of paralogous genes. Such complex relationships require caution in the functional interpretation of the phylogenetic classification of proteins. Nevertheless, about 60% of the original set of 720 COGs [37••] are simple families, with no paralogs or with paralogs from one lineage only, suggesting the possibility of straightforward transfer of functional information from functionally characterized genes from model systems such as *E. coli* and yeast to those from poorly characterized genomes.

The utility of this system of protein classification was tested on several newly sequenced bacterial, archeal and eukaryotic genomes. Interestingly, with the only exception of the minimal genome of *M. genitalium*, the fraction of the proteins that belong to the COGs — ancient families conserved across a wide phylogenetic range — is about the same and very close to 50% for all prokaryotic genomes (Table 1). This is clearly compatible with the previous estimate that about 70% of the proteins encoded in each genome contain ancient conserved regions. The fraction of the proteins included in the COGs is at this time lower, which is evidently due to the requirement for three distant lineages to be included, and to the limited number of species in the first instalment of the COGs. There is little doubt that with new genomes added, the number of COGs will asymptotically approach the total number of ancient conserved regions. By contrast, this fraction is much lower

for eukaryotic genomes, indicating the prevalence of eukaryote-specific families.

Comparison of the new protein sets with the COGs resulted in a number of functional predictions for previously uncharacterized proteins. Even for the *Helicobacter pylori* proteins, most of which show highly significant similarity to homologs from *E. coli* and other bacteria and have been described in considerable detail [8••], predictions were made in more than 100 cases (http://www.ncbi.nlm.nih/COG); function was also predicted for a number of archeal and worm proteins (EV Koonin, RL Tatusov, MY Galperin, unpublished data).

Missing gene families and evolution of metabolic pathways

Comparative analysis of the available complete genomes shows that metabolic diversity generally correlates with genome size. Parasitic bacteria import a variety of metabolites, which allows them to shed genes encoding enzymes for many or even most of the metabolic pathways [1–3, 8••,33,38]. In contrast, all cells have to rely on their own gene products for performing such essential functions as genome expression, replication and repair, and membrane biogenesis and others. These tasks alone require at least about 200 genes [13,37••].

Given complete genome sequences, classification of proteins into orthologous groups provides a convenient way to systematically survey the protein families present or absent in a genome and to identify the metabolic pathways that are likely to be operative in the organism analyzed. When some of the required enzymes cannot be found in the genome, the respective pathways are either not operative, or use other, unrelated, proteins to catalyze the missing steps (see [39]). An example of such an analysis, which included superposition of the phylogenetic patterns derived from the COGs [37••], over the scheme of glycolysis, reveals several interesting trends (Figure 1). Glycolysis includes three reactions that in different species are catalyzed by non-orthologous enzymes, namely phosphofructokinases, aldolases and phosphoglycerate mutases. Interestingly, the second phosphofructokinase in *E. coli*, encoded by the *pfkB* gene, has apparently been recruited from a ubiquitous family of ribokinase-like sugar kinases. The ribokinase COG seems to be an example of a complex family in which the exact orthologous connections are not always easy to trace. In particular, even though PfkB formally belongs to the COG, there seems to be no actual ortholog of it in other genomes. Thus *H. pylori* does not encode a phosphofructokinase at all, although it has genes for other kinases of the ribokinase family and, accordingly, is represented in the respective COG (Figure 1).

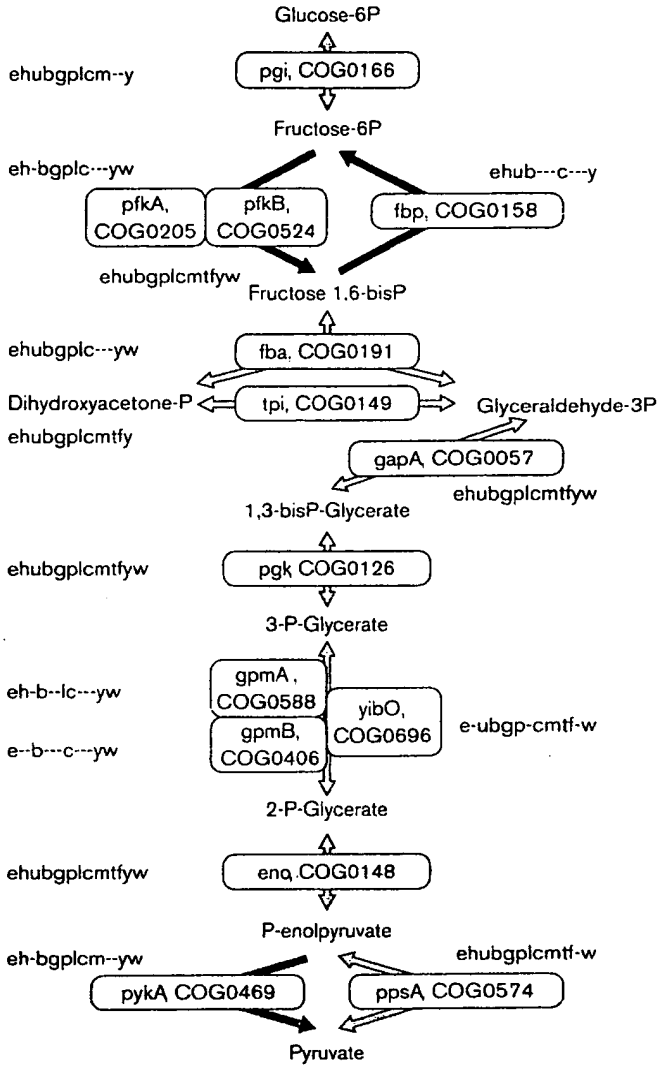
A remarkable case of non-orthologous gene displacement involves two unrelated forms of phosphoglycerate mutase, the 2,3-bisphosphoglycerate (BPG)-dependent and the BPG-independent one. While *H. influenzae* and *Borrelia burgdorferi* encode only the BPG-dependent form, and *H. pylori*, mycoplasmas, and archea encode only the BPG-independent form (see [40]), free-living bacteria such as *E. coli*, *Bacillus subtilis* and *Synechocystis* sp. possess genes coding for both these forms, with two paralogs of the BPG-dependent one (Figure 1). Phosphofructokinase, aldolase and fructose bisphosphatase genes are all missing in the archea (Figure 1), in accordance with the experimental data [41]. This is consistent with the idea that glycolysis originally evolved as a biosynthetic pathway, containing only the lower (tri-carbon) part [42].

Systematic identification of missing links in functional systems in organisms for which complete genome sequences are available is probably the most important application of protein family classification. Conspicuous gaps in the *H. pylori* metabolism became apparent from the COG analysis, suggesting major revisions to the general scheme of the central metabolic pathways in this bacterium (Table 2). In particular, unlike most other bacteria (and all with completely sequenced genomes), *H. pylori* seems to possess neither glycolysis nor the pentose phosphate shunt, the Entner-Doudoroff pathway being the only major route of sugar catabolism. Indeed, sugar fermentation, resulting in intracellular acid production, would be an additional burden on the pH maintenance mechanism in this bacterium, which has to survive in an external pH of 2–3. By contrast, gluconeogenesis, which converts organic acids into sugars required for nucleic acid and peptidoglycan biosynthesis and thus removes H^+ from the cytoplasm, appears to be fully functional in *H. pylori*. For the purpose of energy production, *H. pylori* apparently depends on amino acid fermentation, which causes alkalinization of the cytoplasm and thus relieves part of the problem of pH maintenance. Amino acids and oligopeptides that serve as substrates for this fermentation are produced by gastric proteolysis and transported by readily identifiable permeases.

From genomes and families to superfamilies and folds

Classification systems aimed at the identification of families of orthologs make no attempt to capture the more subtle conserved motifs in proteins, which reflect ancient relationships at the level of superfamilies and frequently are critically important for understanding protein functions and structures [43,44]. Computer methods for the detection of such motifs and delineation of superfamilies have lately progressed significantly through programs such as BLIMPS/MULTIMAT [45], Probe [46], and PSI-BLAST [47••], which combine pairwise sequence comparisons with profile analysis. PSI-BLAST, in particular, has proved to be a powerful tool for the detection of subtle sequence motifs, resulting in the discovery of a number of unsuspected superfamily relationships [47••,48•]. Furthermore, one of the perhaps under-appreciated benefits of the accumulation of genomic sequences is the greatly improved capacity to identify even very subtle sequence similarities due to

Figure 1



Glycolytic enzymes in organisms with completely sequenced genomes. The enzymes are listed under *E. coli* gene names. The COG numbers are as in COG database (www.ncbi.nlm.nih.gov/COG, [37••]) (where available). Shaded arrows indicate reversible reactions, black arrows practically irreversible ones. Phosphoenolpyruvate synthase-catalyzed reaction in the direction of phosphoenolpyruvate hydrolysis has been demonstrated *in vitro*. Phylogenetic patterns are: e, *Escherichia coli*; h, *Haemophilus influenzae*; u, *Helicobacter pylori*; b, *Bacillus subtilis*; g, *Mycoplasma genitalium*; p, *Mycoplasma pneumoniae*; l, *Borrelia burgdorferi*; c, *Synechocystis* sp.; m, *Methanococcus jannaschii*; t, *Methanobacterium thermoautotrophicum*; f, *Archaeoglobus fulgidus*; y, *Saccharomyces cerevisiae*; w, *Caenorhabditis elegans*.

the increasingly uniform population of the protein universe by these relatively unbiased sequence sets, of which the new methods for sequence analysis mentioned above can take advantage [49•].

In the past year, we have seen the identification or significant extension of a number of protein superfamilies; some examples, with the distribution among complete genomes, are shown in Table 3. Most of these superfamilies are universally found in all genomes, with the counts more or less proportional to the total number of genes in the genome. Some expansions are, however, remarkable, such as, for example, urease-related hydrolases and ATP-grasp domains in the archea, and HAD superfamily hydrolases in *E. coli* and *B. subtilis* (Table 3). In certain cases, the phylogenetic distribution of a superfamily immediately suggests major evolutionary events. Thus the BRCT domain is present in a single copy in the DNA ligase of all bacteria (with one additional copy found only in *Synechocystis*), is missing in the archea, and is dramatically expanded in its distribution in the eukaryotes (Table 3). The most obvious interpretation of this distribution is that this domain has entered the eukaryotic world by horizontal gene transfer from bacteria and has undergone exten-

Table 2

Genes and pathways missing in *Helicobacter pylori*.

Enzyme activity	*E. coli* gene	COG number	Status in *H. pylori*	Implications for *H. pylori* metabolism
Phosphofructokinase	*pfkA*	COG0206	Missing	Absence of the two key glycolytic enzymes shows that Embden-Meyerhof pathway is not functional in *H. pylori*. Gluconeogenesis enzymes, bypassing these reactions, fructose bisphosphatase (HP1385) and phosphoenolpyruvate synthase (HP0121), are present in *H. pylori*, allowing it to produce sugars required for peptidoglycan biosynthesis.
	pfkB	COG0525	Present (ribokinase)	
Pyruvate kinase	*pykA*	COG0470	Missing	
	pykF			
6-phosphogluconate dehydrogenase	*gnd*	COG0360	Missing	Pentose phosphate pathway is also not functional. Even though *H. pylori* has a ribose 5-phosphate isomerase encoded by an ortholog of the E. coli *rpiB*, no gene coding for 6-phosphogluconate dehydrogenase could be identified. The only saccharolytic pathway in *H. pylori* appears to be the Entner-Doudoroff pathway.
Ribose 5-phosphate isomerase	*rpiA*	COG0120	Missing	
Lipoate synthase	*lipA*	COG0318	Missing	Pyruvate dehydrogenase complex is absent in *H. pylori*; acetate kinase and phosphotransacetylase are not functional. Pyruvate-ferredoxin oxidoreductase is the only acetyl-CoA-producing enzyme in *H. pylori*.
Lipoate-protein ligase	*lplA*	COG0411	Missing	
	lipB	COG0319	Missing	
Dihydrolipoamide acyltransferase	*aceF*	COG0510	Missing	
Acetate kinase	*ackA*	COG0280	Disrupted by a frameshift	
Phospho-transacetylase	*pta*	COG0278	Disrupted by frameshifts	
Enzymes of purine biosynthesis	*purF*	COG0034	Missing	*De novo* purine biosynthesis is absent in *H. pylori*, and it has to obtain purines from the host. HP1185 appears to be the best candidate for the purine permease, as it is the only *H. pylori* protein, similar to *E. coli* PurP.
	purD	COG0151	Inactivated by mutations	
	purN	COG0299	Missing	
	purT	COG0027	Missing	
	purL_1	COG0046	Missing	
	purL_2	COG0047	Missing	On the other hand, *H. pylori* encodes the enzymes for AMP and GMP synthesis from IMP and their interconversion. Therefore, it can survive on any of these purines.
	purM	COG0150	Missing	
	purK	COG0026	Missing	
	purE	COG0041	Missing	
	purC	COG0152	Missing	
	purH	COG0138	Missing	
	purA	COG0104	Present	
	purB	COG0015	Present	
	guaB	COG0516	Present	
	guaA_1	COG0518	Present	
	guaA_2	COG0519	Present	

sive duplication with divergence in the eukaryotes. The expansion of this domain into a number of eukaryotic proteins involved in cell-cycle control [50••,51] may have been critical for the very establishment of these systems.

With the current acceleration in protein structure determination [22,24], a superfamily identified by sequence comparison more and more frequently extends to include proteins with known 3D structure and/or well-characterized catalytic mechanism (Table 3). Such findings are sometimes most illuminating as they immediately result in the prediction of the structural fold, the structure of the active center, and possibly also the catalytic mechanism for a wide variety of diverse proteins comprising the superfamily. This is illustrated by the recent prediction of the structure and the catalytic amino acid residues for P-ATPases, which remained elusive in spite of a long history of studies, on the basis of the sequence motifs shared with haloacid dehalogenases [52•].

Assignment of the gene products to structural folds and families with maximal attainable precision is arguably one of the foremost tasks of genome analysis after the sequencing phase. The number of structures that have been determined experimentally is negligible for almost all genomes, with the exception of *E. coli* (where it is still rather a small fraction) (Table 1). A database search with a deliberately conservative similarity cut-off already increases the fraction of proteins for which a confident structure prediction is possible to 10–25% [53•] (Table 1). Secondary structure-based threading allows

Table 3

Some recently identified or significantly expanded protein superfamilies.

Superfamily	Enzymes with known 3D structures (PDB codes)	Enzymes with newly predicted properties	Representatives in complete genomes*	References
BRCT (conserved domain in cell cycle checkpoint proteins)	None	DNA polymerase subunit DPB11, terminal deoxynucleotidyltransferases, deoxycytidyl transferase, DNA ligases III and IV, poly(ADP-ribose) polymerase	e-1, h-1, u-1, b-1, g-1, p-1, l-1, c-2, m-0, t-0, f-0, y-9, w-8	[50]
Urease-related metal-dependent hydrolases	Urease (2kauC), phosphotriesterase (1pta), adenosine deaminase (1fkx)	AMP deaminase, adenine deaminase, cytosine deaminase, hydantoinase, dihydroorotase, allantoinase, aminoacylase, imidazolonepropionase, arylphosphatase, chlorohydrolase, formylmethanofuran dehydrogenase	e-13, h-4, u-2, b-6, g-1, p-1, l-1, c-3, m-9, t-10, f-7, y-6, w-9	[55]
Acid phosphatases	Vanadium-containing chloroperoxidase (1vnc)	Phosphatidic acid phosphatase, phosphatidylglycerol phosphatase, diacylglycerol pyrophosphate phosphatase, glucose-6-phosphatase	e-4, h-1, u-3, b-2, g-0, p-0, l-0, c-1, m-3, t-1, f-0, y-7, w-4	[56,57]
ATP-grasp (ATP-dependent C-N and C-S ligases)	Glutathione synthetase (1gsh), D-ala-D-ala ligase (2dln), biotin carboxylase (1bnc), carbamoyl phosphate synthase (1jdb), succinyl-CoA synthetase (1scu)	Phosphoribosylamine-glycine ligase, phosphoribosylglycinamide formyltransferase, phosphoribosylaminoimidazole carboxylase, tubulin-tyrosine ligase, protein S6-glutamate ligase (RimK), malate thiokinase, ATP-citrate lyase	e-10, h-5,u-4, b-12, g-2,p-2, l-1, c-7, m-9, t-8, f-11, y-10, w-15	[58]
HAD (phosphatases and other hydrolases)	L-Haloacid dehalogenase (1jud)	Phosphoserine phosphatase, phosphoglycolate phosphatase, histidinol phosphatase, glycerol-3-phosphatase, sucrose phosphate synthase, phosphomannomutase, P-type cation-transport ATPases	e-10, h-3, u-1, b-11, g-4, p-5, l-2, c-8, m-3, t-4, f-3, y-9, w-6	[52]
DHH (hydrolases)	None	Exopolyphosphatase, 5′-3′ exonuclease	e-1, h-1, u-4, b-6, g-2, p-2, l-2, c-3, m-6, t-3, f-7, y-1, w-0	[59]
Alkaline phosphatase-related metal-dependent hydrolases	Alkaline phosphatase (1alk), N-acetylgalactosamine 4-sulfatase (1fsu), cerebroside sulfatase(1auk)	Phosphopentomutase, 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, streptomycin-6-phosphatase, phosphonoacetate hydrolase, phosphoglycerol transferase, nucleotide pyrophosphatase, steroid sulfatase, aryl- and hexosamine sulfatases	e-15, h-5, u-4, b-8, g-1, p-1, l-0, c-1, m-2, t-2, f-3, y-6, w-18	[40,60]

*Identified by PSI-BLAST [47••] searches of the complete genomes using the conserved motif(s) for each superfamily as a query. Organism abbreviations are as in Figure 1: e, *E. coli*; h, *H. influenzae*; u, *H. pylori*; b, *B. subtilis*; g, *M. genitalium*; p, *M. pneumoniae*; l, *B. burgdorferi*; c, *Synechocystis sp.*; m, *M. jannaschii*; t, *M. thermoautotrophicum*; f, *A. fulgidus*; y, *S. cerevisiae*; w, *C. elegans*.

another relatively small but notable increase in the predictive power [54•] (Table 1). It appears, however, that at this time, the most realistic way to further structure prediction at genome scale is to perform a complete analysis of protein superfamilies as exemplified in Table 3.

Perspective

As far as prokaryotic genomes are concerned, we have already entered the post-genomic era. While surprises certainly wait ahead, there is little doubt that the major protein families are already known or can be deciphered from the available sequences. We have recently seen major progress in methods and procedures for advanced sequence analysis, and a lot of valuable information has been extracted from the genomes. We believe, however, that a major focused effort in genome comparison is still required in order to construct a proper classification of protein families and superfamilies and systematically apply it to the goals of structural and functional prediction. Such an effort will have the potential of creating a basis for a rationally designed, decisive onslaught on structure determination and experimental identification of gene functions using computer predictions as a guide. Hopefully, this research program turns out to be both realistic and efficient.

Ref erences and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269**:496-512.

2. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, **270**:397-403.

3. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1996, **24**:4420-4449.

4. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD *et al.*: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*.** *Science* 1996, **273**:1058-1073.

5. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S *et al.*: **Sequence analysis of the genome of the unicellular *Cyanobacterium synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, **3**:109-136.

6. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes.** *Science* 1996, **274**:546, 563-567.

7. •• Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277**:1453-1474.

The completion of the genome sequence of *E. coli*, one of the classic objects of molecular biology and genetics, certainly has a symbolic significance. More importantly, the enormous amount of information available regarding *E. coli* gene functions can now be used to full potential for inferring functions of homologs in other species. However, the functions of about one half of the *E. coli* genes have not been determined experimentally, and so there is still a lot to learn about *E. coli* itself.

8. •• Tomb J, White O, Kerlavage A, Clayton R, Sutton G, Fleishmann R, Ketchum K, Klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388**:539-547.

The genome sequence of this bacterium is of special interest from several points of view. The genome analysis will have important practical implications as *H. pylori* is the causative agent of peptic ulcers and is believed to infect up to half of the human population. *H. pylori* thrives in a highly acidic environment (pH 2–3); deciphering the mechanisms of acid tolerance from the genome sequence is a most interesting task. Furthermore, *H. pylori* represents an early branching of the proteobacterial lineage, and the comparison of its genome with those of other Proteobacteria such as *E. coli* and *Haemophilus influenzae* will shed light on the evolution of cellular functions in bacteria and mitochondria.

9. •• Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K *et al.*: **Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics.** *J Bacteriol* 1997, **179**:7135-7155.

This second genome of a methanogenic archeon to be sequenced, after *Methanococcus jannaschii*, is of major importance in corroborating trends revealed by the *M. jannaschii* genome analysis [4,34]. Like *M. jannaschii*, there is a sharp divide between the majority of the genes, which appear to have bacterial origin, and a minority (primarily encoding proteins involved in genome replication and expression) of 'eukaryotic' genes. Some other unusual aspects of the *M. jannaschii* genome, however, did not recur in *M. thermoautrophicum*. For example, unlike *M. jannaschii*, *M. thermoautrophicum* encodes a typical set of molecular chaperones such as DnaK and DnaJ and does not encode a unique ATPase family found in *M. jannaschii*.

10. •• Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD *et al.*: **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 1997, **390**:364-370.

The first sequence of a non-methanogenic archeon, and the third complete archeal genome altogether. With 2436 genes, the *A. fulgidus* genome is considerably larger than those of *M. jannaschii* and *M. thermoautrophicum*, in part due to more extensive duplication in some of the gene families. Unlike *M. jannaschii* and *M. thermoautrophicum*, *A. fulgidus* does not seem to encode any inteins. With three genome sequences available, there is for the first time an opportunity for an informative comparative analysis of archeal genomes. Definitive work in this area remains to be done, but it is already clear that the three genomes generally are highly coherent, and also that there are many mysterious conserved families, creating a challenge for further research, both theoretical and experimental.

11. •• Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390**:249-256.

The second classic bacterial model, after *E. coli*, and also the second largest bacterial genome sequenced so far (4100 genes compared with 4288 genes in *E. coli*). With *B. subtilis* adequately representing the Gram-positive lineage (only the minimal genomes of *Mycoplasma* had been available before), we may now have a sampling of the great majority of bacterial gene families. In addition to its value for comparative analysis, *B. subtilis* is most interesting and important in its own right, given, for example the large number of genes in its genome that encode enzymes of secondary metabolite synthesis.

12. •• Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK *et al.*: **Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*.** *Nature* 1997, **390**:580-586.

The first genome representing yet another major division of bacteria, the spirochetes. The genome has a number of unique features, above all a linear chromosome unusual in the bacterial world, and at least 17 linear and circular plasmids that contain about 30% of the genes. Most of the plasmid-borne genes remain quite mysterious, at least after the initial genome analysis.

13. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

14. Koonin EV, Galperin MY: **Prokaryotic genomes: the emerging paradigm of genome-based microbiology.** *Curr Opin Genet Dev* 1997, **7**:757-763.

15. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF: **GenBank.** *Nucleic Acids Res* 1998, **26**:1-7.

16. Dayhoff MO, Barker WC, Hunt LT: **Establishing homologies in protein sequences.** *Methods Enzymol* 1983, **91**:524-545.

17. Barker WC, Garavelli JS, Haft DH, Hunt LT, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LSL, Ledley RS, Mewes HW *et al.*: **The PIR-**

International Protein Sequence Database. *Nucleic Acids Res* 1998, **26**:27-32.

18. Bairoch A, Bucher P, Hofmann K: **The PROSITE database, its status in 1997.** *Nucleic Acids Res* 1997, **25**:217-221.

19. Attwood TK, Beck ME, Flower DR, Scordis P, Selley JN: **The PRINTS protein fingerprint database in its fifth year.** *Nucleic Acids Res* 1998, **26**:306-311.

20. Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, **26**:322-325.

21. Corpet F, Gouzy J, Kahn D: **The ProDom database of protein domain families.** *Nucleic Acids Res* 1998, **26**:325-328.

22. Holm L, Sander C: **Touring protein fold space with Dali/FSSP.** *Nucleic Acids Res* 1998, **26**:318-321.

23. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH – a hierarchic classification of protein domain structures.** *Structure* 1997, **5**:1093-1108.

24. Hubbard TJP, Murzin AG, Brenner SE, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 1997, **25**:236-239.

25. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540.

26. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**:99-113.

27. Fitch WM: **Uses for evolutionary trees.** *Phil Trans R Soc Lond B Biol Sci* 1995, **349**:93-102.

28. Koonin EV, Tatusov RL, Rudd KE: **Sequence similarity analysis of Escherichia coli proteins: functional and evolutionary implications.** *Proc Natl Acad Sci USA* 1995, **92**:11921-11925.

29. Labedan B, Riley M: **Widespread protein sequence similarities: origins of Escherichia coli genes.** *J Bacteriol* 1995, **177**:1585-1588.

30. Labedan B, Riley M: **Gene products of Escherichia coli: sequence comparisons and common ancestries.** *Mol Biol Evol* 1995, **12**:980-987.

31. Riley M, Labedan B: **Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module.** *J Mol Biol* 1997, **268**:857-868.

32. Brenner SE, Hubbard T, Murzin A, Chothia C: **Gene duplications in *H. influenzae*.** *Nature* 1995, **378**:140.

33. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*.** *Curr Biol* 1996, **6**:279-291.

34. • Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25**:619-637.

A detailed comparison of the first available archeal genome (*M. jannaschii*) with bacterial genomes produced a number of novel functional predictions and led to the conclusion that the majority of archeal genes most probably have a bacterial origin. Furthermore, generalizations started to emerge, including the nearly constant fraction of genes containing ancient conserved regions – about 70% in all genomes – and the same major superfamilies of paralogs.

35. Clayton RA, White O, Ketchum KA, Venter JC: **The first genome from the third domain of life.** *Nature* 1997, **387**:459-462.

36. Overbeek R, Larsen N, Smith W, Maltsev N, Selkov E: **Representation of function: the next step.** *Gene* 1997, **191**:GC1-GC9.

37. •• Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.

Comparative analysis of the proteins encoded in seven complete genomes from five major phylogenetic lineages and elucidation of consistent patterns of sequence similarities resulted in the delineation of 720 clusters of orthologous groups (COGs). Each COG consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, allowing transfer of functional information from one member to an entire COG. This automatically makes possible a number of functional predictions, especially for poorly characterized genomes. The evolving system of COGs comprises a framework for functional and evolutionary genome analysis; it is accessible through the World Wide Web (http://ncbi.nlm.nih.gov/COG).

38. Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R: **Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*.** *Nucleic Acids Res* 1997, **25**:701-712.

39. Koonin EV, Mushegian AR, Bork P: **Non-orthologous gene displacement.** *Trends Genet* 1996, **12**:334-336.

40. Galperin MY, Bairoch A, Koonin EV: **A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases.** *Protein Sci* 1998, **7**:in press.

41. Danson MJ: **Central metabolism of the archaea.** In *The Biochemistry of Archaea (Archaebacteria)*. Edited by Kates M, Kushner DJ, Matheson AT. Amsterdam: Elsevier; 1993:1-24.

42. Romano AH, Conway T: **Evolution of carbohydrate metabolic pathways.** *Res Microbiol* 1996, **147**:448-455.

43. Bork P, Koonin EV: **Protein sequence motifs.** *Curr Opin Struct Biol* 1996, **6**:366-376.

44. Bork P, Gibson TJ: **Applying motif and profile searches.** *Methods Enzymol* 1996, **266**:162-184.

45. Henikoff S, Henikoff JG: **Embedding strategies for effective use of information from multiple sequence alignments.** *Protein Sci* 1997, **6**:698-705.

46. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE: **Extracting protein alignment models from the sequence database.** *Nucleic Acids Res* 1997, **25**:1665-1677.

47. •• Altschul SF, Madden TL, Schaffer AA, Zhang J, Zheng Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST - A new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

A major revamp of BLAST, which is definitely the most popular current method for database search. The key innovations are: first, the program now makes gapped alignments, with appropriately modified statistics, which results in significant increase of sensitivity; and second, the associated program PSI (Position-Specific Iterating)-BLAST makes a position-specific weight matrix (profile) out of the first pass results and iterates searches with this profile until no new sequences with similarity scores above a defined cut-off are detected. This appears to be the most powerful existing method for detection of subtle similarities between protein sequences and delineation of protein superfamilies.

48. • Mushegian AR, Bassett DE Jr, Boguski MS, Bork P, Koonin EV: **Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs.** *Proc Natl Acad Sci USA* 1997, **94**:5831-5836.

Sequence analysis of the proteins encoded by 70 positionally cloned human disease genes showed that most of them have orthologs with the same domain architecture in the nematode, but domain rearrangements are prevalent in yeast and bacterial homologs. This is one of the first demonstrations of the utility of PSI-BLAST for the delineation of large protein superfamilies. In particular, this method was used for the identification of a conserved ATPase domain present in the repair protein MutL (one of the colon cancer gene products in humans), histidine kinases, molecular chaperones of the HSP90 family and type II DNA topoisomerases; the 3D structure for the latter was already available, defining the fold for the whole superfamily.

49. • Bork P, Koonin EV: **Predicting functions from protein sequences: where are the bottlenecks?** *Nature Genet* 1998, **18**:313-318.

An attempt to analyze the reasons why it is so common that functionally and phylogenetically important relationships between sequences are not detected in original analysis (particularly in the framework of genome projects) but are readily identified in subsequent, more detailed studies. It appears that the major bottlenecks include inadequate filtering for noise in sequence data (for example low-complexity sequences and very common domains) and insufficient cross-talk between different types of information.

50. •• Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV: **A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins.** *FASEB J* 1997, **11**:68-76.

A complete description of the BRCT domain that had been originally found in BRCA1 protein and several other proteins implicated in cell cycle checkpoint. In this work, the superfamily has been extended to include a distinct version of the BRCT domain detected in bacterial DNA ligases, the large subunits of eukaryotic replication factor C, and poly(ADP-ribose) polymerases. The expansion of the BRCT domain in eukaryotes may be one of the key events in the evolution of cell-cycle control.

51. Callebaut I, Mornon JP: **From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair.** *FEBS Lett* 1997, **400**:25-30.

52. Aravind L, Galperin MY, Koonin EV: **The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold.** *Trends Biochem Sci* 1998, **23**:127-129.
This paper is an example of the application of sequence profile analysis to the prediction of the 3D fold and the catalytic residues in a critically important enzyme, P-ATPase, which has defied crystallization attempts and remained poorly characterized in spite of intense effort.

53. Frishman D, Mewes HW: **PEDANTic genome analysis.** *Trends Genet* 1997, **13**:415-416.
This paper describes a very convenient Worldwide Web site compiling results of automatic analysis of all available complete genomes. The Pedant WWW site (http://pedant.mips.biochem.mpg.de/frishman/pedant.html) is arguably one of the best entry points to comparative genomics but it has to be kept in mind that it is only the first level, crude analysis that is presented here.

54. Fischer D, Eisenberg D: **Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*.** *Proc Natl Acad Sci USA* 1997, **94**:11929-11934.
One of the first systematic attempts to predict the 3D structures of proteins starting from a complete genome. The utility of sequence–structure threading is demonstrated but it also becomes clear that such methods at best result in a rather small, incremental improvement over state-of-the-art sequence comparisons. Although the fraction of the proteins with a predictable fold is only 22% of the gene products, the authors predict by extrapolation that it should be possible to assign folds to most soluble proteins within a decade.

55. Holm L, Sander C: **An evolutionary treasure: unification of a broad set of amidohydrolases related to urease.** *Proteins* 1997, **28**:72-82.
A valuable example of a combination of detailed sequence analysis with structure–structure comparisons resulting in the characterization of a vast protein superfamily.

56. Stukey J, Carman GM: **Identification of a novel phosphatase sequence motif.** *Protein Sci* 1997, **6**:469-472.

57. Neuwald AF: **An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases.** *Protein Sci* 1997, **6**:1764-1767.

58. Galperin MY, Koonin EV: **A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity.** *Protein Sci* 1997, **6**:2639-2643.

59. Aravind L, Koonin EV: **A novel family of predicted phosphoesterases includes *Drosophila* prune protein and bacterial RecJ exonuclease.** *Trends Biochem Sci* 1998, **23**:17-19.

60. Bond CS, Clements PR, Ashby SJ, Collyer CA, Harrop SJ, Hopwood JJ, Guss JM: **Structure of a human lysosomal sulfatase.** *Structure* 1997, **5**:277-289.